

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2000-276308

(P2000-276308A)

(43)公開日 平成12年10月6日(2000.10.6)

(51)Int.Cl. ⁷	識別記号	F I	テ-マ-ト*(参考)
G 0 6 F 3/06	5 4 0	G 0 6 F 3/06	5 4 0 5 B 0 1 8
	3 0 5		3 0 5 A 5 B 0 6 5
			3 0 5 C
			3 0 5 F
12/16	3 1 0	12/16	3 1 0 Q
審査請求 未請求 請求項の数 8 O L (全 13 頁) 最終頁に続く			

(21)出願番号 特願平11-85205

(22)出願日 平成11年3月29日(1999.3.29)

(71)出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72)発明者 阿部 秀一

東京都青梅市末広町2丁目9番地 株式会

社東芝青梅工場内

(74)代理人 100083161

弁理士 外川 英明

Fターム(参考) 5B018 GA02 GA06 HA04 HA13 HA14

HA35 MA12

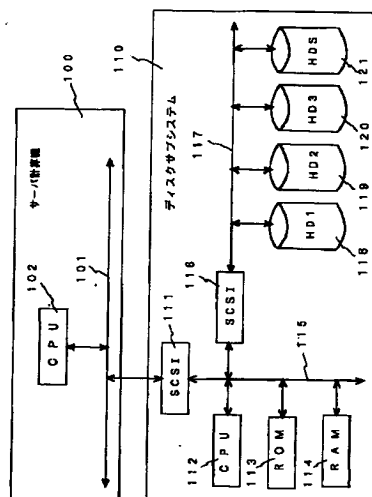
5B065 BA01 CA30 EA02 EA12 EA24

(54)【発明の名称】 ディスクサブシステム及びディスクサブシステムにおけるデータ復元方法

(57)【要約】

【課題】本発明は、障害が発生したデータの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してディスクサブシステムが使用不能になることを防止するディスクサブシステムを提供することを目的とする。

【解決手段】RAIDを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を設けたことを特徴とする。



【特許請求の範囲】

【請求項1】予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、

障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を具備したことを特徴とするディスクサブシステム。

【請求項2】前記データ復元手段は、演算によりデータを復元する際、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記予備の磁気ディスク装置に登録すること特徴とする請求項1記載のディスクサブシステム。

【請求項3】予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、

障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記予備の磁気ディスク装置に登録するデータ復元手段を具備したことを特徴とするディスクサブシステム。

【請求項4】複数の磁気ディスク装置でRAIDが構成され、一つの磁気ディスク装置に障害が発生した場合にはその障害が発生した磁気ディスク装置を新しい磁気ディスク装置に交換するディスクサブシステムにおいて、前記障害の発生した磁気ディスク装置のデータを前記交換した新しい磁気ディスク装置に復元する際に、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記交換した新しい磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記交換した新しい磁気ディスク装置に登録するデータ復元手段を具備したことを特徴とするディスクサブシステム。

【請求項5】予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、

障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納することを特徴とするディスクサブシステムにおけるデータ復元方法。

【請求項6】前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記予備の磁気ディスク装置に登録すること特徴とする請求項5記載のディスクサブシステムにおけるデータ復元方法。

【請求項7】予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、

障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記予備の磁気ディスク装置に登録することを特徴とするディスクサブシステムにおけるデータ復元方法。

【請求項8】複数の磁気ディスク装置でRAIDが構成され、一つの磁気ディスク装置に障害が発生した場合にはその障害が発生した磁気ディスク装置を新しい磁気ディスク装置に交換するディスクサブシステムにおいて、前記障害の発生した磁気ディスク装置のデータを前記交換した新しい磁気ディスク装置に復元する際に、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記交換した新しい磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記交換した新しい磁気ディスク装置に登録することを特徴とするディスクサブシステムにおけるデータ復元方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、PCサーバ等のサーバ計算機において、高信頼性が要求される外部記憶装置として使用されるRAID技術を使用したディスクサブシステムの改良に関する。特に信頼性を更に高くするRAID技術を使用したディスクサブシステムに関する。

【0002】

【従来の技術】RAID (Redundant Arrays of Inexpensive Disks) 技術は、David A. Patterson, Garth A. Gribson, Randy H. Katsにより1987年に発表された論文「A Case for Redundant Arrays of Inexpensive Disks」で紹介されて実用化されている。

【0003】そしてこのRAID技術は、一般的にレベル1からレベル5に分類され、高信頼性が要求されるPCサーバ等のサーバ計算機のディスクサブシステムに使用されている。

【0004】レベル5のRAID技術を使用したディスクサブシステムの一例を図8を用いてその概要を説明する。図8において、ディスクサブシステムは、3台の磁気ディスク装置DR1、DR2、DR3と一台の予備磁気ディスク装置DRSとから構成されている。この図においては、サーバ計算機やRAIDコントローラの図示を省略している。

【0005】サーバ計算機が管理し磁気ディスク装置に記録しているデータを所定単位のブロックに分割し、これを3台の磁気ディスク装置DR1、DR2、DR3に分散して記録する。この際に、冗長データとしてパリティを作成し、このパリティも分散して記録する。

【0006】このパリティは、次のような方法で生成される。パリティP1は、データDaとデータDbとの排他的論理和演算を行うことで生成される。同様にパリティP2はデータDcとデータDdとの排他的論理和演算で、パリティP3はデータDeとデータDfとの排他的論理和演算で、パリティP4はデータDgとデータDhとの排他的論理和演算で生成される。

【0007】このように構成されたディスクサブシステムにおいて、サーバ計算機からデータDbのリード要求があった場合に、磁気ディスク装置DR2のデータDbが記録されているメディア（磁気ディスク）の記録面の障害又は磁気ディスク装置DR2全体の障害のため、データDbの読み出しができないときには、再度データリードを試みてリトライを実施し、それでも読み出しができない場合には、RAIDコントローラがデータDaとパリティP1との排他的論理和演算を行いデータDbを復元してサーバ計算機にリードデータとして出力してサーバ計算機にはディスクサブシステムに障害が発生し

たことを認識させずに対応している。

【0008】このように障害が発生した場合、RAIDコントローラは障害が発生した磁気ディスク装置DR2をRAID構成から切り離して磁気ディスク装置DR1とDR3の2台の構成による縮退状態にする。

【0009】この縮退状態で更に別な磁気ディスク装置に故障が発生すると、ディスクサブシステムからのデータの読み出しができなくなり、RAIDコントローラはサーバ計算機に対してディスクサブシステムに障害が発生したと通知することになる。

【0010】このように縮退状態で更に別な磁気ディスク装置に故障が発生してディスクサブシステムに障害が発生させないように、RAIDを正常な状態に回復する必要がある。

【0011】この回復処理は、上記縮退状態において予備の磁気ディスク装置DRSに切り離した磁気ディスク装置DR2のデータを復元することで行われる。この予備の磁気ディスク装置DRSに切り離した磁気ディスク装置DR2のデータを復元するには、磁気ディスク装置DR1とDR3のそれぞれに対応するブロックのデータどうしの排他的論理和演算を行うことで実現できる。例えばデータDeを復元するには、パリティP3とデータDfとの排他的論理和演算で復元できる。

【0012】また、予備の磁気ディスク装置を持たないディスクサブシステムの回復処理では、障害の発生した磁気ディスク装置を別の正常な磁気ディスク装置と交換し、この交換した磁気ディスク装置にデータを上記と同様に排他的論理和演算により復元する。

【0013】

【発明が解決しようとする課題】このようなRAID技術を使用したディスクサブシステムにおいては、次のような問題点があった。障害が発生しRAID構成から切り離した磁気ディスク装置のデータを予備の又は交換した磁気ディスク装置に復元する際に、上記切り離した磁気ディスク装置の障害の内容がメディア（磁気ディスク）の記録面の場合には、障害の発生していないメディアからはそこに記録されているデータの読み出しが可能にもかかわらず、読み出しが可能なデータも含めて全てのデータを排他的論理和演算により求めるため、データの復元処理に大幅な時間を要していた。

【0014】また、このようなデータの復元処理の際に、排他的論理和演算のためにRAID構成している障害の発生していない磁気ディスク装置からデータを読み出す際に、障害が発生してデータの読み出しができない場合がある。このとき、この時点でデータの復元処理は不可能になるばかりでなく、ディスクサブシステム自体が使用不可になってしまっていた。このようにデータの復元処理の際に障害が見つかるデータは、普段アクセスがされていないあまり重要なデータではないと推測される。このようなあまり重要でないデータが読み出せない

ことに起因して、障害の発生した磁気ディスク装置のデータの復元ができなくなるとともに、ディスクサブシステム自体が使用不能になることは大きな問題であった。

【0015】本発明は、これら従来の問題点を解決するためになされたもので、障害が発生した磁気ディスク装置のデータの復元処理を高速化できるディスクサブシステムを提供することを目的とする。

【0016】また、本発明は、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してRAIDが再構成できなくなり、ディスクサブシステムが使用不能になることを防止するディスクサブシステムを提供することを目的とする。

【0017】

【課題を解決するための手段】本発明は、予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を設けたことを特徴とする。

【0018】このような構成によれば、データの復元処理を高速化できる。また、本発明は、予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記障害の発生した磁気ディスク装置から読み出せるデータは、そのデータを前記予備の磁気ディスク装置に格納し、前記障害の発生した磁気ディスク装置から読み出せないデータは、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するデータ復元手段を設け、更に前記演算によりデータを復元する際、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記予備の磁気ディスク装置に登録することとを特徴とする。

【0019】このような構成によれば、データの復元処理を高速化できるとともに障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置か

ら読み出せないことに起因してRAIDが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【0020】また、本発明は、予備の磁気ディスク装置と複数の磁気ディスク装置とでRAIDを構成したディスクサブシステムにおいて、障害の発生した磁気ディスク装置のデータを前記予備の磁気ディスク装置に復元する際に、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記予備の磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記予備の磁気ディスク装置に登録するデータ復元手段を設けたことを特徴とする。

【0021】このような構成によれば、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してRAIDが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【0022】更に、本発明は、複数の磁気ディスク装置でRAIDが構成され、一つの磁気ディスク装置に障害が発生した場合にはその障害が発生した磁気ディスク装置を新しい磁気ディスク装置に交換するディスクサブシステムにおいて、前記障害の発生した磁気ディスク装置のデータを前記交換した新しい磁気ディスク装置に復元する際に、前記RAIDを構成する複数の磁気ディスク装置の内、障害が発生していない磁気ディスク装置に格納されているデータから演算により復元し、この復元したデータを前記交換した新しい磁気ディスク装置に格納するとともに、前記演算によりデータを復元するとき、その演算に用いられるデータが前記障害が発生していない磁気ディスク装置から読み出せない場合には、その復元対象のデータがECCエラーであると前記交換した新しい磁気ディスク装置に登録するデータ復元手段を設けたことを特徴とする。

【0023】このような構成によれば、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必要なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してRAIDが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【0024】

【発明の実施の形態】以下、図面を参照して本発明の第1の実施形態について説明する。図1は本発明の第1の実施形態に関わるシステムの概略構成を説明する図である。図1には、PCサーバなどのサーバ計算機100と

ディスクサブシステム110とからなる計算機システムの概略構成を図示している。

【0025】サーバ計算機100にはシステムバス101が設けられており、このシステムバス101にCPU102が接続されている。また、システムバス101にはSCSIインタフェース(I/F)111を介してディスクサブシステム110が接続されている。尚、SCSIは、Small Computer System Interfaceの省略語である。

【0026】ディスクサブシステム110は、CPU112、ROM113、RAM114、SCSIインタフェース(I/F)111とSCSIインタフェース(I/F)116がバス115を介して相互に接続されているRAID制御部とSCSIバス117を介してSCSIインタフェース(I/F)116に接続された4台の磁気ディスク装置HD1、HD2、HD3、HDSとから構成されている。

【0027】CPU112は、ROM113に格納されたファームウェアに基づいて、ディスクサブシステム110の全体を制御する。RAM114には、サーバ計算機100からディスクサブシステム110に対するデータのリード又はライトの命令が出された際のアドレス情報を実際の磁気ディスク装置に記録されているアドレス情報に変換する変換テーブルが設けられている。また、RAM114は、CPU112の動作に伴う各種データが保存される。磁気ディスク装置HD1と磁気ディスク装置HD2と磁気ディスク装置HD3とでレベル5のRAIDを構成している。そして、磁気ディスク装置HDSは、予備の磁気ディスク装置として設けられている。

【0028】図2には、サーバ計算機100から見える論理的にディスクサブシステム110に記録されているデータの配置の構成と実際にRAIDを構成する個々の磁気ディスク装置に記録されているデータとの関係を図示している。

【0029】サーバ計算機100から見て論理的にディスクサブシステム110に記録されているデータ200は、所定サイズの6つのブロックに分割されたデータD1～D6に分割されている。この6つのデータD1～D6は、図示のように3台の磁気ディスク装置HD1、HD2、HD3に記録されている。磁気ディスク装置HD1には、パリティデータP1、データD3、データD5とが記録されている。同様に、磁気ディスク装置HD2には、データD2、パリティデータP2、データD6とが記録されている。また、磁気ディスク装置HD3には、データD2、データD4、パリティデータP3とが記録されている。磁気ディスク装置HDSは、予備の磁気ディスク装置であるため、現時点では何も記録されていない。パリティP1は、データD1とデータD2との排他的論理和演算により求められたデータである。同様にパリティP2は、データD3とデータD4との排他的

論理和演算により求められたデータである。また、パリティP3は、データD5とデータD6との排他的論理和演算により求められたデータである。

【0030】以下、図3、図4、図5に示すフローチャートに基づいて、本発明の第1の実施形態におけるデータの復元動作について説明する。尚、データの復元動作中においても、並行してサーバ計算機100からのディスクサブシステム110へのデータの書き込み及び読み出しの命令を受け付け、その処理を行うものとする。復元動作中にサーバ計算機100からデータの書き込み又は読み出しの命令を受け付けた場合の処理は、以下の動作説明の中で説明する。

【0031】以下、レベル5のRAIDを構成している図1に図示したディスクサブシステム110において、メディアの障害により磁気ディスク装置HD3からデータD2の読み出しができなくなり、RAID制御装置のCPU112が磁気ディスク装置HD3をRAID構成から切り離して縮退状態とし、予備の磁気ディスク装置HDSに磁気ディスク装置HD3に格納されていたデータを復元する場合のCPU112の制御に基づく動作を説明する。

【0032】まず、磁気ディスク装置HD3に格納されていたデータを予備の磁気ディスク装置HDSに復元するに際して、CPU112はRAM114上に図6に示すようなデータ変更/復元マップ300を作成する(ステップS1)。このデータ変更/復元マップ300は、データ301、状態ビット302、復元ビット303の3行から構成されており、データを復元する磁気ディスク装置HD3の復元動作中におけるデータの変更及び復元の状態を管理するために設けられている。従って、データ301としては、D2、D4、P3が登録される。また、この初期時には状態ビット302と復元ビット303は、「0」に設定される。復元ビット303は、以下に説明するデータの復元処理が終了した状態をデータ(ブロック)毎に表すものである。このビットが「1」に設定されているということは、その対応するデータが予備の磁気ディスク装置HDSに復元されたことを意味する。また、状態ビット302は、この復元動作中にサーバ計算機100からの書き込み命令により、対応するブロックのデータが変更された場合にセットされる。

【0033】次に変数Bを「1」を設定する(ステップS2)。この変数Bは、磁気ディスク装置HD3の何番目のブロックのデータを復元動作するかを示すものである。次にフラグ情報である変数SETを「1」を設定する(ステップS3)。この変数SETが「1」にセットされている場合は、CPU112が磁気ディスク装置HD3のブロック単位のデータを予備の磁気ディスク装置HDSに復元動作中であることを示す。次にB番目のブロックのデータの復元処理を行う(ステップS4)。即ち、1番目のブロック(データD1が格納されているブ

ロック)のデータ(データD1)の復元処理を行う。

【0034】この復元処理の詳細は、図7に示すフローチャートを用いて説明する。まず、該当する1番目のブロックの状態ビットに「1」が設定されているかどうかをデータ変更/復元マップ300を参照してチェックする(ステップS50)。このチェックの結果、状態ビットに「1」が設定されている場合には、ステップS52へ進む。現時点では、状態ビットに「0」が設定されているため、ステップS51に進む。

【0035】次に、1番目のブロックのデータD2が障害の発生した磁気ディスク装置HD3から読み出せるかをチェックする(ステップS51)。読み出せる場合には、磁気ディスク装置HD3からデータD2を読み出し、予備の磁気ディスク装置HDSの対応するブロックにコピーして復元する(ステップS53)。そして、復元動作を終了する。もし、読み出せない場合には、ステップS52に進む。

【0036】次に1番目のブロックのデータD2が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティP1とデータD1とが、それぞれ磁気ディスク装置HD1及び磁気ディスク装置HD2から読み出せるかをチェックする(ステップS52)。読み出せる場合には、ステップS54へ進み、パリティP1とデータD1との排他的論理和演算を行い、データD2を復元し、予備の磁気ディスク装置HDSの対応するブロックにコピーする(ステップS54)。そして、復元動作を終了する。また、読み出せない場合には、ステップS55へ進む。

【0037】ステップS55では、現在復元処理をしているブロックのデータが、障害の発生した磁気ディスク装置HD3から読み出せず、更に排他的論理和演算によっても復元できない状態のため、予備の磁気ディスク装置HDSの対応するブロックにECC(Error Checking and Correcting)エラーが発生したとして、予備の磁気ディスク装置HDSのメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにECCエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行い、図3のステップS4のデータの復元処理が終了する。

【0038】次に図3のステップS5に進む。ステップS5では、ステップS4においてデータD2が復元されたので、図6に示すデータD2に対応する復元ビットを「1」にセットする。

【0039】次にステップS6では、変数Bを+1してインクリメントする。続いてステップS7において、変数SETを「0」を設定する。この変数SETを「0」にセットすることで、CPU112が磁気ディスク装置HD3のブロック単位のデータを予備の磁気ディスク

装置HDSに復元する動作が終了したことを意味する。

【0040】続いて処理はステップS8に進み、既に復元したブロックのデータがサーバ計算機100からのデータの書き込み命令により変更されていないかをチェックする。具体的には、図6に示したデータ変更/復元マップ300において、復元ビット303が「1」に設定されているデータの内、状態ビット302が「1」に設定されているデータの有無をチェックする。このチェックの結果、復元ビット303と状態ビット302の双方が「1」に設定されているデータがある場合には、そのデータが既にデータが復元されたもので、且つその後のサーバ計算機100からのデータの書き込み命令により変更されているものであると判断されステップS10へ進む。また、復元ビット303と状態ビット302の双方が「1」に設定されているデータがない場合には、ステップS9に進む。以下、ステップS9に進む場合の動作を説明し、ステップS10へ進む場合の動作は、後に説明する。

【0041】ステップS9では、全てのブロックのデータが復元されたかを変数Bの値を見ることでチェックする。全てのブロックのデータが復元された場合には、全ての復元処理が終了する。全てのブロックのデータが復元されていない場合には、ステップS3へ戻り以上説明した動作を繰り返す。

【0042】次に以上説明したステップS3からS9の処理を繰り返している間にサーバ計算機100からディスクサブシステム110へのデータの書き込み又は読み出しの命令をCPU112がSCSIインタフェース111を介して受理した場合には、以上説明したデータの復元処理動作を中断して、そのデータの書き込み又は読み出し処理を実行し、その後復元処理動作を再開する。

【0043】まず、CPU112がサーバ計算機100からデータの読み出し命令を受理した場合について説明する。RAIDを構成する正常な磁気ディスク装置からのデータの読み出しであれば、そのデータを読み出す。もし、障害が発生しRAID構成から切り離れた磁気ディスク装置HD3からのデータの読み出しの場合には、もしそのデータが磁気ディスク装置HD3から読み出せる場合には、そのまま読み出し、読み出せない場合には、他の正常な磁気ディスク装置から読み出した対応するブロックのデータの排他的論理和演算により求めたデータを読み出しデータとしてサーバ計算機100からへ転送する。ただし、このように説明したいずれの読み出しの場合においても、メディアの障害が原因によりデータが読み出せないときには、上述したようにECCエラーとして登録し、サーバ計算機100からにその旨を通知する。

【0044】次にCPU112がサーバ計算機100からデータの書き込み命令を受理した場合について説明する。まず、上記変数SETが「0」であることを確認す

る。もし、上記変数SETが「1」の場合には、「0」になるまで、その処理を待機させる。変数SETが「0」であることを確認できた場合には、以下の処理を行う。RAIDを構成する正常な磁気ディスク装置に記録されているデータの書き込みであれば、そのまま書き込み処理を実施する。この際、パリティの変更をすることは説明するまでもない。

【0045】また、障害が発生しRAID構成から切り離した磁気ディスク装置HD3へのデータの書き込みの場合、例えばデータD4の書き込みの場合には、次のように行う。まず、この新しい書き込みデータD4と磁気ディスク装置HD1に書き込まれているデータD3とで排他的論理和演算を行い得られたパリティデータを新しいパリティP2として磁気ディスク装置HD2に書き込む。そして、新しい書き込みデータD4を実際に書き込む代わりに図6に示したデータ変更/復元マップ300のデータD4に対応する状態ビット302を「1」にセットする。この様な処理をすることで書き込み処理を実施する。

【0046】次に図4のステップS8において、復元ビット303と状態ビット302の双方が「1」に設定されており、既にそのデータが復元されたもので、且つその後サーバ計算機100からのデータの書き込み命令により変更されているデータがあると判断されステップS10へ進む場合の動作を説明する。以下の説明ではデータ変更/復元マップ300が図4に示した状態になっている場合を例にして説明する。

【0047】ステップS10では、フラグ情報である変数SETを「1」を設定する。続いて、ステップS11では、ステップS8において復元ビット303と状態ビット302の双方が「1」に設定されていると判断されたデータD2の復元処理を再度実行する。この復元処理は、図7で示したルーチンに基づき行われるもので、既に説明したので、ここではその説明を省略する。

【0048】次にステップS12で図6に示したデータ変更/復元マップ300におけるデータD2に対応する状態ビット302を「0」に設定する。続いてステップS13において、変数SETを「0」を設定しステップS8に戻る。この様に復元ビット303と状態ビット302の双方が「1」に設定されているデータが無くなるまで、ステップS8→S10→S11→S12→S13の処理を続ける。

【0049】次に本発明の第2の実施形態について説明する。第1の実施形態との違いは、図7に示した復元処理の動作が異なり、他の動作は、第1の実施形態と同一である。

【0050】第2の実施形態における復元処理の動作を図8に示したフローチャートを用いて説明する。動作説明の前提としては、第1の実施形態と同様にレベル5のRAIDを構成している図1に図示したディスクサブシ

ステム110において、メディアの障害により磁気ディスク装置HD3からデータD2の読み出しができなくなり、RAID制御装置のCPU112が磁気ディスク装置HD3をRAID構成から切り離して縮退状態とし、予備の磁気ディスク装置HDSに磁気ディスク装置HD3に格納されていたデータを復元する場合のCPU112の制御に基づく動作を説明する。

【0051】まず、1番目のブロックのデータD2が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティP1とデータD1とが、それぞれ磁気ディスク装置HD1及び磁気ディスク装置HD2から読み出せるかをチェックする（ステップS60）。読み出せる場合には、ステップS61へ進み、パリティP1とデータD1との排他的論理和演算を行い、データD2を復元し、予備の磁気ディスク装置HDSの対応するブロックにコピーする（ステップS61）。そして、復元動作を終了する。

【0052】また、読み出せない場合には、ステップS62へ進む。ステップS62では、1番目のブロックのデータD2が障害の発生した磁気ディスク装置HD3から読み出せるかをチェックする（ステップS62）。読み出せる場合には、磁気ディスク装置HD3からデータD2を読み出し、予備の磁気ディスク装置HDSの対応するブロックにコピーして復元する（ステップS63）。そして、復元動作を終了する。

【0053】もし、読み出せない場合には、ステップS64に進む。ステップS64では、現在復元処理をしているブロックのデータが、排他的論理和演算によっても復元できず、更に障害の発生した磁気ディスク装置HD3から読み出せない状態のため、予備の磁気ディスク装置HDSの対応するブロックにECCエラーが発生したとして、予備の磁気ディスク装置HDSのメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにECCエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行い、図3のステップS4のデータの復元処理が終了する。

【0054】次に本発明の第3の実施形態について説明をする。第1の実施形態との違いは、図7に示した復元処理の動作が異なることと、ディスクサブシステム110に予備の磁気ディスク装置を持たず、RAIDを構成する磁気ディスク装置に障害が発生した場合には、その磁気ディスク装置を新しい正常な磁気ディスク装置と交換することである。その他の動作は、第1の実施形態と同一である。

【0055】第3の実施形態における復元処理の動作を図9に示したフローチャートを用いて説明する。動作説明の前提としては、第1の実施形態と同様にレベル5のRAIDを構成している図1に図示したディスクサブシ

る。もし、上記変数SETが「1」の場合には、「0」になるまで、その処理を待機させる。変数SETが

「0」であることを確認できた場合には、以下の処理を行う。RAIDを構成する正常な磁気ディスク装置に記録されているデータの書き込みであれば、そのまま書き込み処理を実施する。この際、パリティの変更をするとは説明するまでもない。

【0045】また、障害が発生しRAID構成から切り離れた磁気ディスク装置HD3へのデータの書き込みの場合、例えばデータD4の書き込みの場合には、次のように行う。まず、この新しい書き込みデータD4と磁気ディスク装置HD1に書き込まれているデータD3とで排他的論理和演算を行い得られたパリティデータを新しいパリティP2として磁気ディスク装置HD2に書き込む。そして、新しい書き込みデータD4を実際に書き込む代わりに図6に示したデータ変更/復元マップ300のデータD4に対応する状態ビット302を「1」にセットする。この様な処理をすることで書き込み処理を実施する。

【0046】次に図4のステップS8において、復元ビット303と状態ビット302の双方が「1」に設定されており、既にそのデータが復元されたもので、且つその後サーバ計算機100からのデータの書き込み命令により変更されているデータがあると判断されステップS10へ進む場合の動作を説明する。以下の説明ではデータ変更/復元マップ300が図4に示した状態になっている場合を例にして説明する。

【0047】ステップS10では、フラグ情報である変数SETを「1」を設定する。続いて、ステップS11では、ステップS8において復元ビット303と状態ビット302の双方が「1」に設定されていると判断されたデータD2の復元処理を再度実行する。この復元処理は、図7で示したルーチンに基づき行われるもので、既に説明したので、ここではその説明を省略する。

【0048】次にステップS12で図6に示したデータ変更/復元マップ300におけるデータD2に対応する状態ビット302を「0」に設定する。続いてステップS13において、変数SETを「0」を設定しステップS8に戻る。この様に復元ビット303と状態ビット302の双方が「1」に設定されているデータが無くなるまで、ステップS8→S10→S11→S12→S13の処理を続ける。

【0049】次に本発明の第2の実施形態について説明する。第1の実施形態との違いは、図7に示した復元処理の動作が異なり、他の動作は、第1の実施形態と同一である。

【0050】第2の実施形態における復元処理の動作を図8に示したフローチャートを用いて説明する。動作説明の前提としては、第1の実施形態と同様にレベル5のRAIDを構成している図1に図示したディスクサブシ

ステム110において、メディアの障害により磁気ディスク装置HD3からデータD2の読み出しができなくなり、RAID制御装置のCPU112が磁気ディスク装置HD3をRAID構成から切り離して縮退状態とし、予備の磁気ディスク装置HDSに磁気ディスク装置HD3に格納されていたデータを復元する場合のCPU112の制御に基づく動作を説明する。

【0051】まず、1番目のブロックのデータD2が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティP1とデータD1とが、それぞれ磁気ディスク装置HD1及び磁気ディスク装置HD2から読み出せるかをチェックする(ステップS60)。読み出せる場合には、ステップS61へ進み、パリティP1とデータD1との排他的論理和演算を行い、データD2を復元し、予備の磁気ディスク装置HDSの対応するブロックにコピーする(ステップS61)。そして、復元動作を終了する。

【0052】また、読み出せない場合には、ステップS62へ進む。ステップS62では、1番目のブロックのデータD2が障害の発生した磁気ディスク装置HD3から読み出せるかをチェックする(ステップS62)。読み出せる場合には、磁気ディスク装置HD3からデータD2を読み出し、予備の磁気ディスク装置HDSの対応するブロックにコピーして復元する(ステップS63)。そして、復元動作を終了する。

【0053】もし、読み出せない場合には、ステップS64に進む。ステップS64では、現在復元処理をしているブロックのデータが、排他的論理和演算によっても復元できず、更に障害の発生した磁気ディスク装置HD3から読み出せない状態のため、予備の磁気ディスク装置HDSの対応するブロックにECCエラーが発生したとして、予備の磁気ディスク装置HDSのメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにECCエラーが発生したことのフラグを登録する。この様な一連の動作でデータの復元処理を行う。図3のステップS4のデータの復元処理が終了する。

【0054】次に本発明の第3の実施形態について説明する。第1の実施形態との違いは、図7に示した復元処理の動作が異なることと、ディスクサブシステム110に予備の磁気ディスク装置を持たず、RAIDを構成する磁気ディスク装置に障害が発生した場合には、その磁気ディスク装置を新しい正常な磁気ディスク装置と交換することである。その他の動作は、第1の実施形態と同一である。

【0055】第3の実施形態における復元処理の動作を図9に示したフローチャートを用いて説明する。動作説明の前提としては、第1の実施形態と同様にレベル5のRAIDを構成している図1に図示したディスクサブシ

システム110において、メディアの障害により磁気ディスク装置HD3からデータD2の読み出しができなくなり、磁気ディスク装置HD3を新しい正常な磁気ディスク装置と交換し、この交換した新しい磁気ディスク装置（図示せず）に磁気ディスク装置HD3に格納されていたデータを復元する場合のCPU112の制御に基づく動作を説明する。

【0056】まず、1番目のブロックのデータD2が排他的論理和演算で復元できるかをチェックする。換言すると、この排他的論理和演算をするためのデータであるパリティP1とデータD1とが、それぞれ磁気ディスク装置HD1及び磁気ディスク装置HD2から読み出せるかをチェックする（ステップS70）。読み出せる場合には、ステップS71へ進み、パリティP1とデータD1との排他的論理和演算を行い、データD2を復元し、交換した新しい磁気ディスク装置の対応するブロックにコピーする（ステップS71）。そして、復元動作を終了する。

【0057】また、読み出せない場合には、ステップS72へ進む。ステップS72では、現在復元処理をしているブロックのデータが、排他的論理和演算によっても復元できないため、交換した新しい磁気ディスク装置の対応するブロックにECCエラーが発生したとして、交換した新しい磁気ディスク装置のメディアの特定エリアに設定した図示しないエラー状態マップに該当するブロックにECCエラーが発生したことのフラグを登録する。

この様な一連の動作でデータの復元処理を行い、図3のステップS4のデータの復元処理が終了する。

【0058】以上の説明では、復元処理をしているブロックのデータが、復元できないため、ECCエラーが発生したとして、データの復元をしている磁気ディスク装置にECCエラーが発生したことのフラグを登録すると説明した。しかし、その後、その復元できなかったデータのデータがサーバ計算機100により新たにデータが書き込まれた場合には、そのフラグの登録を削除すればよい。

【0059】尚、このようにECCエラーが発生したことのフラグが登録されるデータは、従来技術の説明で述べたように、普段アクセスがされていないあまり重要なデータではないと推測される。従って、データが復元できなくてRAIDの再構成ができなくなり、ディスクサブシステム自体が使用不能になることに比べて、復元できないデータにECCエラーが発生したと登録する処理の方が、システム全体から見れば、より良い対策である。

【0060】

【発明の効果】以上説明した通り、本発明によれば、デ

ータの復元処理を高速化できるとともに、障害が発生した磁気ディスク装置のデータの復元処理の際に、データの復元処理に必用なデータが障害の発生していない磁気ディスク装置から読み出せないことに起因してRAIDが再構成できなくなり、ディスクサブシステムが使用不能になることを防止できる。

【図面の簡単な説明】

【図1】本発明の第1の実施形態に関わるシステムの概略構成を示す図である。

【図2】サーバ計算機から見える論理的にディスクサブシステムに記録されているデータの配置の構成と実際にRAIDを構成する個々の磁気ディスク装置に記録されているデータとの関係を示す図。

【図3】本発明の第1の実施形態におけるデータの復元動作を説明するためのフローチャート図。

【図4】本発明の第1の実施形態におけるデータの復元動作を説明するためのフローチャート図。

【図5】本発明の第1の実施形態におけるデータの復元動作を説明するためのフローチャート図。

【図6】データ変更/復元マップを示す図。

【図7】図3におけるデータ復元処理の詳細を説明するためのフローチャート図。

【図8】本発明の第2の実施形態におけるデータ復元処理の詳細を説明するためのフローチャート図。

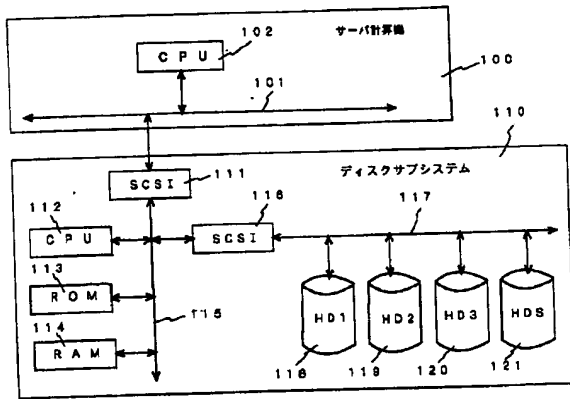
【図9】本発明の第3の実施形態におけるデータ復元処理の詳細を説明するためのフローチャート図。

【図10】従来技術を説明するための、サーバ計算機から見える論理的にディスクサブシステムに記録されているデータの配置の構成と実際にRAIDを構成する個々の磁気ディスク装置に記録されているデータとの関係を示す図。

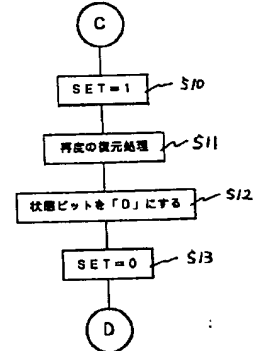
【符号の説明】

100…サーバ計算機
110…ディスクサブシステム
111…SCSIインタフェース
112…CPU
113…ROM
114…RAM
115…バス
116…SCSIインタフェース
117…SCSIバス
118…磁気ディスク装置HD1
119…磁気ディスク装置HD2
120…磁気ディスク装置HD3
121…磁気ディスク装置HD4

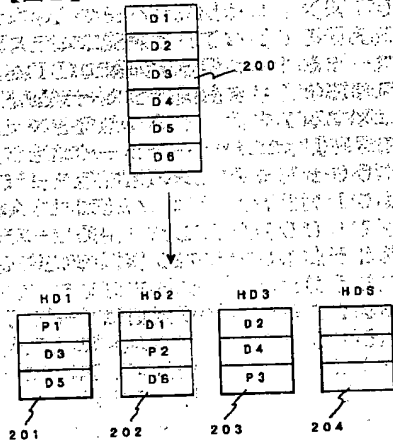
【図1】



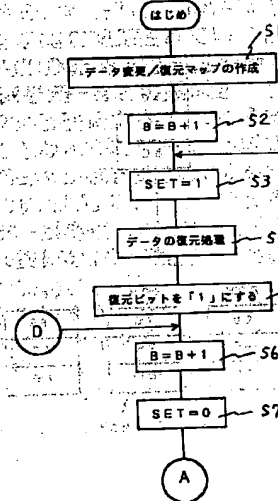
【図5】



【図2】



【図3】

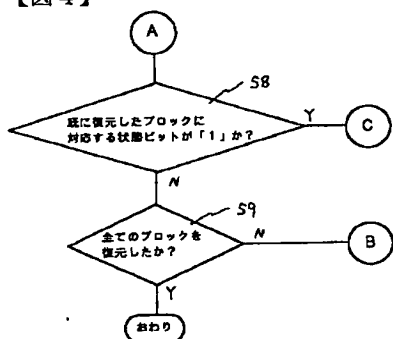


【図6】

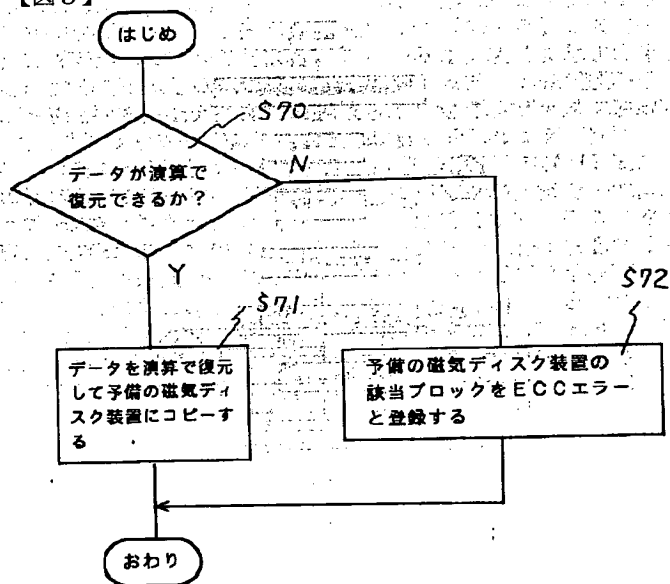
300

301	データ	D2	D4	P3
302	状態ビット	1	0	0
303	復元ビット	1	1	0

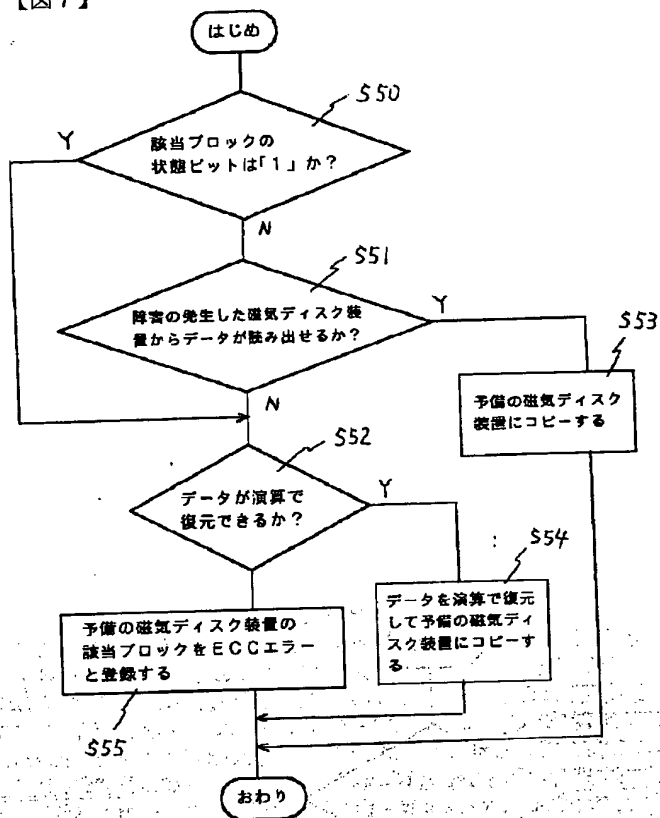
【図4】



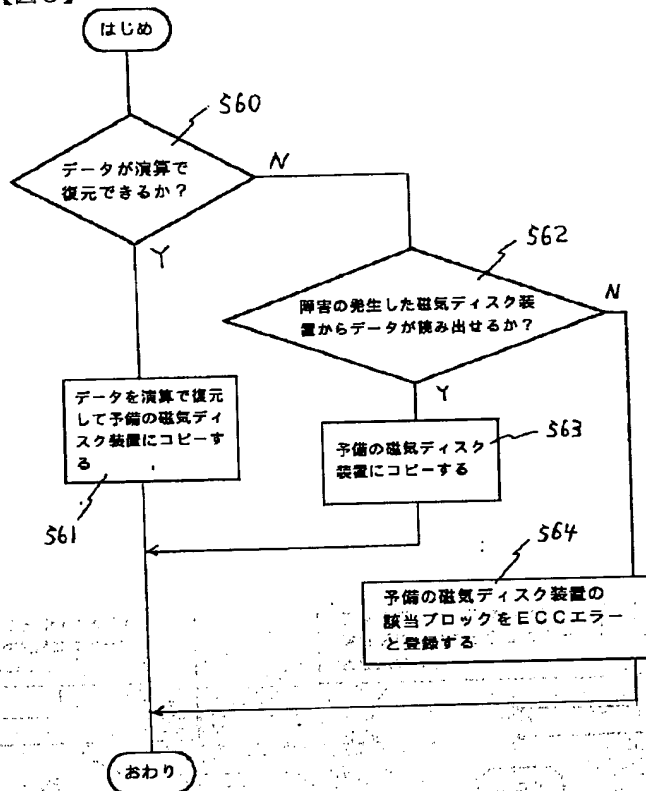
【図9】



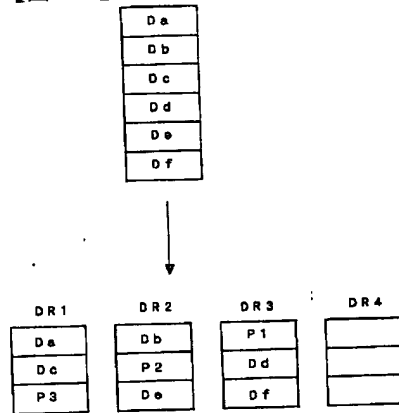
【図7】



【図8】



【図10】



フロントページの続き

(51) Int. Cl. 7 識別記号
G 0 6 F 12/16 3 2 0

(51) Int. Cl. 7 識別記号 テーマコード* (参考) R 1 6
G 0 6 F 12/16 3 3 2 0 L (G 0 6 F 12/16)